IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 July 2023, accepted 25 August 2023, date of publication 30 August 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3310428

RESEARCH ARTICLE

Multi-Modal CNN Features Fusion for Emotion Recognition: A Modified Xception Model

H. M. SHAHZAD[1,2], **SOHAIL MASOOD BHATTI**[1,2], **ARFAN JAFFAR**[1,2], **MUHAMMAD RASHID**[3], **AND SHEERAZ AKRAM**[1,2,4]

[1]Faculty of Computer Science and Information Technology, Superior University, Lahore 55150, Pakistan
[2]Intelligent Data Visual Computing Research (IDVCR), Lahore 55150, Pakistan
[3]Department of Computer Science, National University of Technology, Islamabad 45000, Pakistan
[4]Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 12571, Saudi Arabia

Corresponding author: H. M. Shahzad (shahzad.dar@superior.edu.pk)

ABSTRACT Facial expression recognition (FER) is advancing human-computer interaction, especially, today, where facial masks are commonly worn due to the COVID-19 pandemic. Traditional unimodal techniques for facial expression recognition may be ineffective under these circumstances. To address this challenge, multimodal approaches that incorporate data from various modalities, such as voice expressions, have emerged as a promising solution. This paper proposed a novel multimodal methodology based on deep learning to recognize facial expressions under masked conditions effectively. The approach utilized two standard datasets, M-LFW-F and CREMA-D, to capture facial and vocal emotional expressions. A multimodal neural network was then trained using fusion techniques, outperforming conventional unimodal methods in facial expression recognition. Experimental evaluations demonstrated that the proposed approach achieved an accuracy of 79.81%, a significant improvement over the 68.81% accuracy attained by the unimodal technique. These results highlight the superior performance of the proposed approach in facial expression recognition under masked conditions.

INDEX TERMS Deep learning, multimodal fusion techniques, neural network, facial expression under the mask.

I. INTRODUCTION

As artificial intelligence continues to advance rapidly, there has been a noticeable shift towards emphasizing the importance of emotional intelligence. Nowadays, more and more individuals emphasize developing and understanding their emotional intelligence, in tandem with technological progressions [1], [2]. Facial expressions are incredibly informative as they provide a window into a person's emotions, character, and intentions. This knowledge has countless practical applications, including identifying health concerns and spotting fraudulent behavior. Additionally, analyzing facial expressions can help us understand a person's temperament, psychological state, and even the likelihood of engaging in criminal activity. It is truly unique how much we can learn by observing a person's face [3].

Emotion recognition is the most accurate when the whole face is visible. During a global pandemic (Covid-19), people are encouraged to wear masks that cover their mouths and noses; it may be hard to recognize emotions [4].

In expression analysis, researchers have made significant strides in identifying the association between distinct facial regions and specific emotional categories. Specifically, it has been observed that the mouth plays a pivotal role in accurately recognizing emotions such as happiness, surprise, sadness, disgust, and anger [5]. According to research [6], [7], it has been found that the success of emotion recognition in the lower and upper facial is dependent on the specific emotion being considered. The lower part of the face is better at recognizing happy expressions, but the upper portion of the face seems to be crucial in recognizing afraid and sad faces [8], [9], [10]. Overall, researchers concur that facial expression recognition is reduced when facial features are hidden. The mouth and eyes are the most significant facial features for

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

interpreting facial expressions. The results indicate that visual information from the lower parts of the face plays a vital role in interpreting facial expressions.

Facial expression recognition is the identification of human emotions based on facial expressions. Numerous factors, such as facial characteristics, lighting, and head attitude, make it a tough challenge. The study reveals that people can better interpret the emotions of those whose faces are visible than those whose faces are obscured. The study aimed to determine if wearing a face mask affects how well emotion recognition works. When the nose and mouth were covered, it became more difficult to understand people's emotions [11].

According to recent research, individuals wearing masks may be more challenging to identify with precision due to the concealment of emotions such as happiness, sadness, and anger. Studies have shown that participants are less accurate at identifying emotions when the person is wearing a mask [12], leading to a decrease in overall from 69.9% for unmasked faces to 48.2% for masked target faces [6].

Research has found that deep learning-based facial expression recognition approaches outperform traditional machine learning algorithms in accuracy. The proposed method involves using a neural network with multiple layers to analyze facial images and learn expression characteristics. The approach was tested on the FERET database, which did not include masks, and the results show that it achieves better recognition accuracy than other methods [13]. Furthermore, deep learning has succeeded in single-domain datasets, and current research combines multimodal inputs. Multimodality involves a system that utilizes multiple sensors to extract and combine important information. This results in a more comprehensive representation and better performance when the objects are hidden or partially occluded [14], [15]. Instead of relying only on the masked dataset, we can incorporate other relevant information that complements the masked dataset. The term ''information modality'' describes relying on one or more senses other than sight and hearing to gather and evaluate data before deciding. Compared to single-modal methods, multimodal approaches yield better results [16], [17]. The performance of multimodal deep learning systems, which utilize many modalities, including text, picture, audio, and video, is better than that of individual modalities (i.e., unimodal) systems [18].

Multimodal fusion technology is a powerful tool that can help us make more accurate decisions by processing data from various information sources. It has been used successfully in many fields, including health monitoring, environmental monitoring, machine diagnostics, and aerospace engineering [19].

II. LITERATURE REVIEW

Researchers are increasingly turning to multimodal approaches [20], which involve the integration of diverse data sources. This method utilizes multiple data streams and is gradually gaining prominence within the research community. Several data sources are combined to form a more accurate prediction. It may involve combining information from various sources, such as multiple sensors, different time periods, or different locations. The model can better capture uncertainty and make more accurate forecasts by incorporating data from various sources.

Experts in the field are exploring various methods to enhance CNN accuracy when dealing with intricate data. One such approach involves incorporating multiple modes to improve accuracy. By incorporating additional hidden characteristics, the accuracy of this technique can be further enhanced. This is a promising area of research for improving CNN accuracy with complex data [21].

The researcher explored a multimodal facial recognition approach incorporating low-level facial key point features and a high-level self-learning feature. The experiments revealed that this proposed method outperformed single-modal features, demonstrating its effectiveness in recognizing faces [22]. In a similar work, the Multichannel Convolutional Neural Network (MCCNN) method was proposed and tested on the FER dataset. The results show that the proposed MCCNN works better than the traditional CNN-based architectures. Moreover, the performance of multimodal deep learning systems, which utilize many modalities, including text, picture, audio, and video, is better than that of individual modalities (i.e., unimodal) systems [23].

In general, when it comes to multimodal fusion [24], [25], [26], [27], different fusion levels can be achieved depending on the objectives of the fusion. These levels typically include data level, feature level, and decision level. The data level is about integrating similar sensor data, while the feature level integrates heterogeneous sensor data. Finally, the decision level helps to obtain the final result through multi-source data fusion.

Since it is less probable to identify facial features when wearing a mask effectively—since some emotions and facial expressions may be hidden and difficult to read—a multimodal method is presented in this work to recognize facial expressions while wearing a mask. In recent years, multimodal facial expression feature extraction has emerged as a new area of research and has garnered significant attention. Aiming at the problem of single-source facial expression, many researchers presented multimodal techniques to improve the system even if the objects are hided or partially covered. This paper proposes a multimodal approach to identifying facial expressions while wearing a mask.

III. DATASETS

To develop our multi-modal neural network and fine-tune our model, we utilized the M-LFW-FER dataset [28] for the masked dataset and the CREMA-D [29] dataset for the voice expression dataset.

The details of both datasets are in the next subsection.

A. MLF-W-FER DATASET

The authors [28] have introduced the M-LFW-FER dataset, a masked facial expression recognition dataset created by

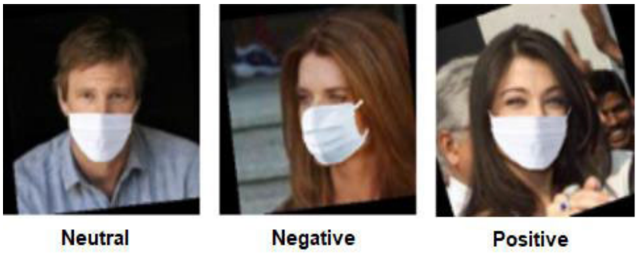

FIGURE 1. The M-LFW-FER dataset.

TABLE 1. Statistic of MLF-FER-W database.

MLF-W-FER	Positive	Negative	Neutral	Total Sample Sets
Training	5194	766	3377	9307
Testing	644	95	416	1155

manually labelling the M-LFW-FER dataset as mentioned in Figure 1. It contains 9307 images of faces wearing masks with labels for neutral, positive, and negative expressions. The testing dataset has 1,155 pictures, with statistics shown in Table 1.

B. CREMA-D

The CREMA-D (Crowd-Sourced Emotional Multimodal Actors Database) dataset [29] trains and evaluates emotion recognition models. The dataset contains audio recordings of seven emotions: anger, contempt, fear, pleasure, sorrow, surprise, and neutrality. CREMA-D has 7,442 images from 91 performers and actresses of various ages and ethnic backgrounds [30]. These speech expressions are publicly accessible in Waveform Audio File (WAV) format.

The CREMA-D dataset is a valuable resource for training and evaluating models for recognizing emotions in different situations. It includes a wide range of emotions, such as anger, disgust, fear, happiness, surprise, and neutrality. The comprehensive dataset ensures that models can recognize and differentiate between various emotional states. Researchers and developers can use this diverse dataset to measure the effectiveness of their approaches, making it a valuable tool for advancing emotion recognition studies. With audio recordings from 91 performers and actresses of different ages and ethnic backgrounds, this dataset offers a diverse range of samples to ensure that trained models are more generalized and applicable to real-world scenarios where emotions may be expressed differently across different individuals and demographics.

IV. PROPOSED MODAL

A. PRE-PROCESSING AND DATA AUGMENTATION

The M-LFW-FER dataset only has three categories, namely positive, neutral, and negative, whereas the CREMA-D dataset has more categories, seven to be precise, including happy, neutral, disgust, sad, surprise, and fear. However, we focused on only three voice expressions - happy, angry, and neutral for our purpose. We ensured that the categories' names were consistent across both models. As a result, we had to change the terms of two categories in the CREMA-D dataset, namely, happy to positive and angry to negative. However, we left the name of the neutral category unchanged.

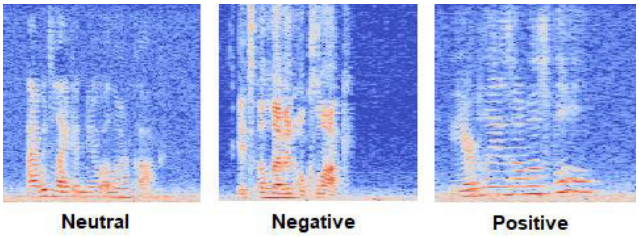

FIGURE 2. Spectrogram conversion for the CREM-D dataset.

CREMA-D recordings are in WAV format and transformed into spectrogram images for voice dataset preprocessing. Spectrograms [31] show voice frequency changes over time, offering insights into pitch variations. The LIBROSA library creates these mel-frequency spectrogram files, which are subsequently cropped to remove unnecessary data and highlight the relevant information as mentioned in Figure 2. A spectrogram is a tool used to analyze the frequencies present in a sound wave, with an amplitude representing signal intensity and colors indicating different frequencies. It is commonly employed in signal, speaker, and speech recognition, and has been extensively used in speech analysis. Research studies have identified the spectrogram as an effective technique for evaluating various acoustic characteristics of speech [32], [33].

When using a multimodal approach, it is crucial to ensure that both the M-LFW-FER (masked) and the CREMA-D dataset have balanced expression classifications. To achieve this balance, we applied an augmentation technique to the CREMA-D dataset to increase the number of voice expressions and match the number of expressions in the M-LFW-FER dataset. We achieved this balance by applying an augmentation technique to the CREMA-D dataset, increasing the number of voice expressions to match the number of expressions in the M-LFW-FER dataset. Our augmentation technique introduces random rotations to the voice spectrograms, with a probabilit of 70%, limited to a maximum of 10 degrees in clockwise and counterclockwise directions. This technique enhances the variability of spectrogram representations, promoting robustness in emotion recognition tasks. Additionally, we explored the use of zoom augmentation with a probability of 50%, allowing for random zooming in or out of voice spectrograms within the range of 1.1 to 1.5. This technique simulates various perspectives of the audio data, contributing to improved generalization capabilities of emotion recognition models.

B. XCEPTION ARCHITECTURE

The Xception architecture, also known as Extreme Inception, is a deep learning model distinguished by its utilization of 36 convolutional layers as mentioned in Figure 3. To address computational efficiency, it utilizes a technique

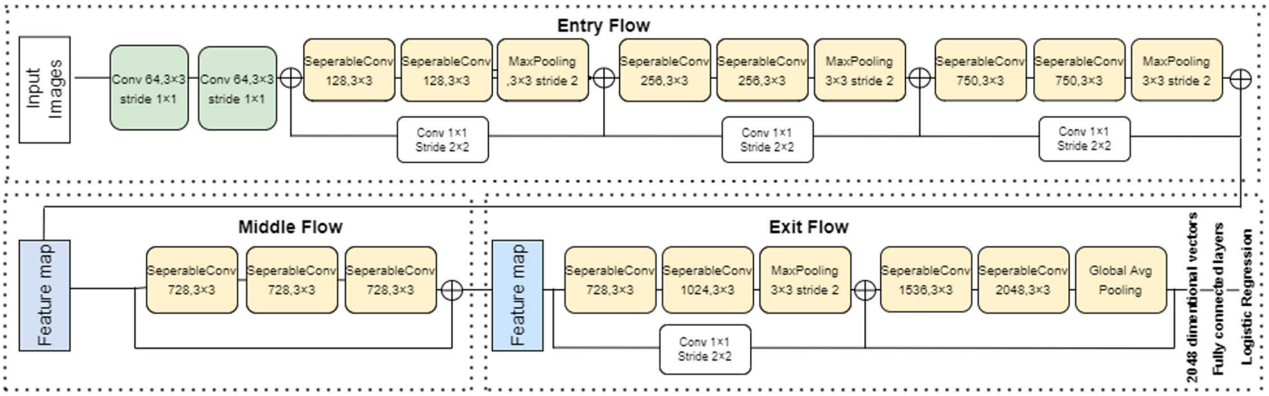

FIGURE 3. Basic architecture of the Xception model.

called depth-wise separable convolutions. This method performs spatial convolutions individually for each channel, significantly reducing computational requirements [34]. Depthwise separable convolutions are comprised of two layers: depthwise spatial convolution and pointwise 1 × 1 convolution. The remarkable aspect of Xception is that it achieves a speed enhancement of 9 times compared to regular convolutions while delivering similar levels of accuracy. This architecture was made explicitly for the ImageNet dataset, an extensive collection of 1.2 million images representing 1,000 distinct classes [35], [36].

C. PROPOSED UNIMODAL MODEL ARCHITECTURE

We applied several modifications to the model and proposed to improve its performance. We combined Convolution2D and Separable Conv2D with L2 regularization (0.0001) in the entry flow, followed by three Separable Conv2D layers in the middle flow, each with 512 kernel filters. The exit flow included two Separable Conv2D layers with 750 kernel filters, followed by a Max pool layer. The feature extraction technique involved converting the resulting feature maps into a single vector for further processing in the neural network, as shown in Figure 4.

In addition, we incorporated a neural network component consisting of three hidden layers, with 3048 neurons and a dropout ratio of 0.3 in the first layer, 2048 neurons and a dropout of 0.2 in the second layer, and 1028 neurons and a dropout of 0.1 in the last layer. To address the issue of overfitting and improve the model's generalization ability, we employed regularization techniques such as dropout and ridge regression. These techniques prevent the model from relying excessively on specific features and optimize its performance. The Adam optimizer was used with a learning rate of 0.0001, and this modified Xception architecture was utilized across all three techniques: data level fusion, feature level fusion, and decision level fusion.

D. INCORPORATING FUSION METHODS IN THE PROPOSED ARCHITECTURE

In our pursuit of enhanced emotion detection, we employed the capabilities of the Xception architecture through a range of fusion methodologies. These encompassed data level, feature level, and decision level fusion techniques, all meticulously orchestrated to elevate accuracy. Central to this innovation was integrating extracted features from the subject's concealed facial and voice expressions. This integration, achieved through a dedicated feature fusion layer, catalyzed amplifying accuracy, forging a harmonious synthesis of multi-modal insights. By capitalizing on the strengths of the Xception architecture and the strategic fusion of diverse inputs, we aimed to push the boundaries of emotion detection, transcending traditional methods and charting a course toward more precise and comprehensive results.

Additionally, regularization techniques such as dropout and ridge regression have been integrated into the model to prevent overfitting. Dropout [37] removes certain features during training, promoting the discovery of multiple independent representations and enhancing generalization, while ridge regression [38] reduces model complexity and prevents overfitting by adding a penalty term to the loss function that encourages small weights.

The Xception architecture's approach is advantageous over traditional architectures like VGG16 or Inception due to its significant reduction in parameters and computational complexity. This reduction ultimately leads to faster training and inference times, making it the optimal choice for various applications. The main difference lies in the use of depthwise separable convolutions, which allows for better efficiency and parameter reduction, ultimately leading to improved performance in terms of speed and sometimes even accuracy.

1) EARLY FUSION TECHNIQUE FOR DATA-LEVEL FUSION

In the first experiment, we merged the data from different modalities and fused their features before passing them through the neural network, as illustrated in Figure 5. The early fusion technique is an intelligent way to deal with the challenges of combining masked facial expressions and voice expressions. This methodology provides a gateway to interweave a spectrum of modalities at the feature level, with the potential to augment the holistic efficiency of the system. This mix of information is used right at the beginning of the architecture before passing to the convolution of the modified

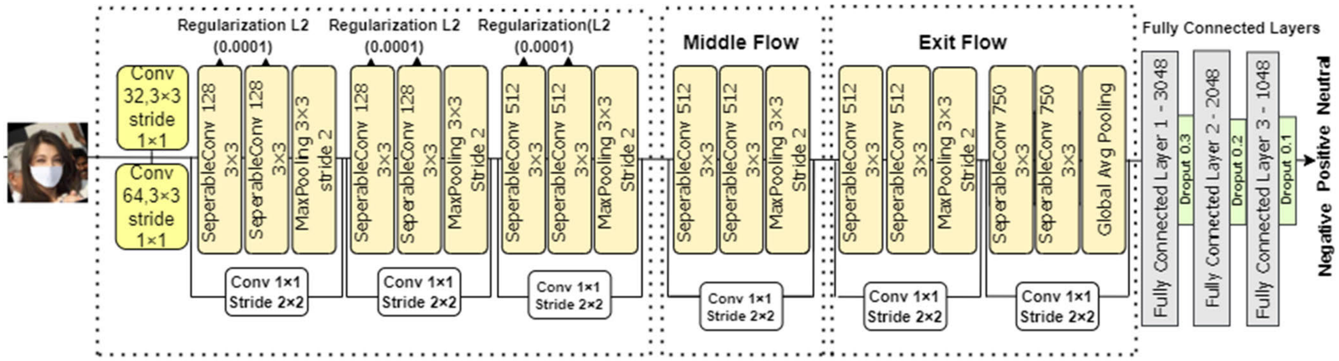

FIGURE 4. Flow chart of the unimodal architecture.

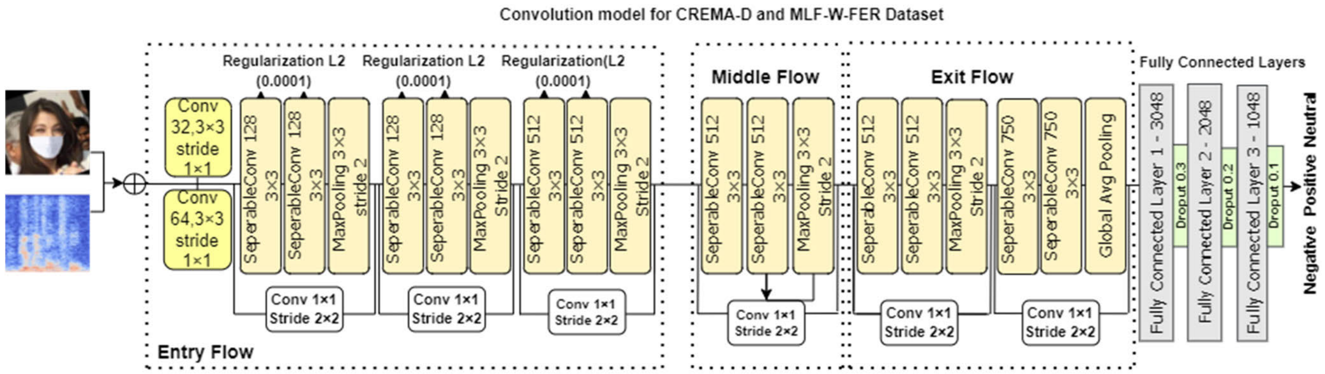

FIGURE 5. Flow chart of the data level fusion using multimodal architecture.

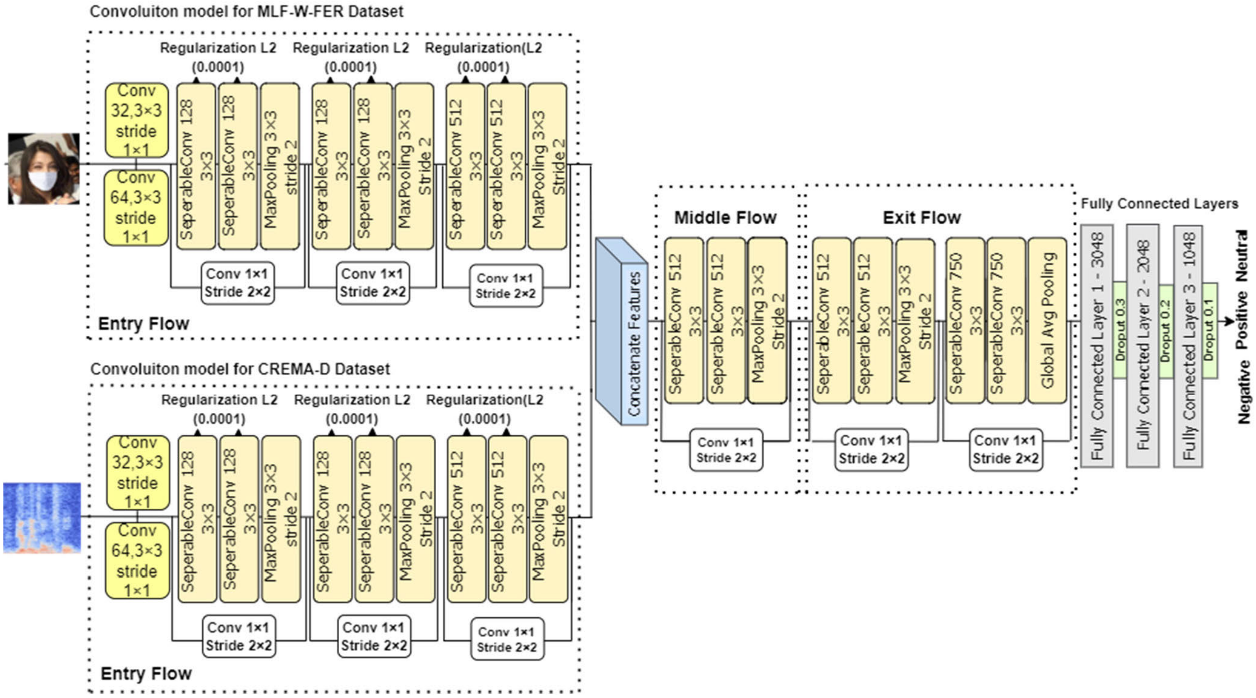

FIGURE 6. Flow chart of the feature level fusion using multimodal architecture.

Exception architecture. This approach proved incredibly effective in enabling us to integrate diverse modalities at the feature level, ultimately leading to enhanced overall system performance.

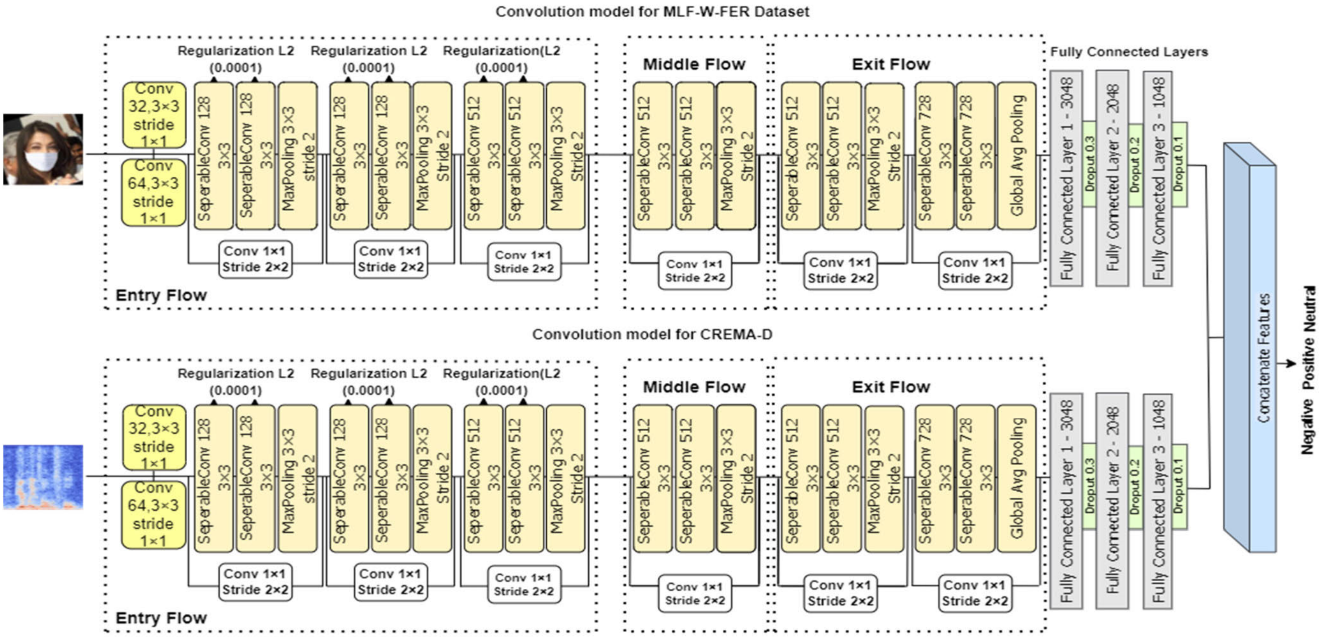

FIGURE 7. Flow chart of the decision level fusion using multimodal architecture.

2) MIDDLE FUSION TECHNIQUE FOR FEATURE-LEVEL FUSION

In the second experiment, we pursued an approach by merging features and combining feature maps from different channels into a single unified feature map at the convolution layer following the entry flow. Subsequently, we processed this consolidated feature map in the middle of the network before directing it to the exit flow for final computations. The visual representation of this process can be seen in Figure 6. Through this technique, we aimed to enhance the integration and transformation of data as it moved through the network's different layers. Combining facial expressions under masks with vocal expressions is a powerful approach that can significantly enhance various human-computer interaction applications, emotion recognition, and multimodal data analysis. Moreover, by leveraging the connection between visual and auditory cues, this technique has the potential to provide highly accurate and comprehensive insights into human emotional states and intentions.

3) LATE FUSION TECHNIQUE FOR DECISION-LEVEL FUSION

In our last experiment, we utilized a technique called late fusion. This approach involved converging the outputs from two networks at a more advanced processing stage. We employed two separate models tailored to process their respective datasets to accomplish this task. Subsequently, we combined the outcomes generated by each model by averaging the results generated by each network within their final fully connected layers, as shown in Figure 7. This approach would be highly effective in achieving our research objectives.

TABLE 2. Accuracies for unimodal on MLF-W-FER Dataset.

Model	Accuracy
AlexNet [1]	55.00
VGG-16 [1]	56.73
VGG-19 [28]	56.79
ResNet-50 [1]	56.80
MobileNet [28]	66.41
MMCFF (Proposed)	**68.80**

V. RESULTS AND DISCUSSION

Our proposed architecture has demonstrated a remarkable level of performance, surpassing other techniques by a significant margin. It achieved an impressive unimodal accuracy of 68.80% on the MLF-W-FER dataset, as illustrated in Table 2. The comparison table highlights the accuracy of alternative models employing unimodal techniques on this dataset. Specifically, the MobileNet achieved an accuracy of 66.41% for masked facial expressions, while the ResNet 50 model attained an accuracy of 56.80%.

Furthermore, we conducted experiments using the CREMA-D dataset to evaluate the accuracy of our proposed architecture. As shown in Table 3, our suggested architecture achieved the highest level of precision at 72.39%. This outperformed other models, with VGG-16 achieving an accuracy of 60.17% and VGG-16 attaining 57.23%.

To enhance the recognition efficiency of masked facial expressions, we conducted experiments in a multimodal architecture and applied various fusion techniques to further improve the model.

Our study involved three experiments that utilized early fusion, middle fusion, and late fusion techniques. In the first

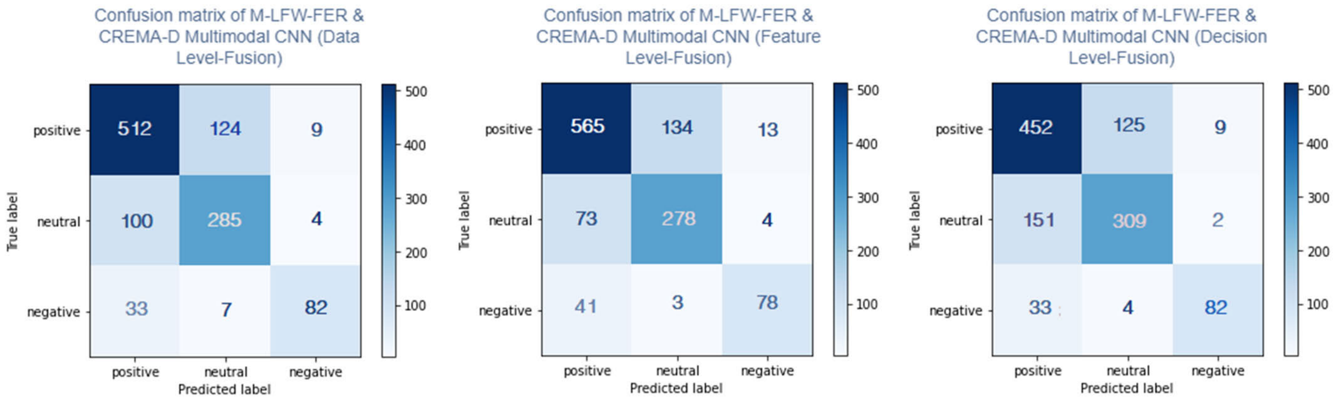

FIGURE 8. Confusion matrix of multimodal model confusion techniques.

TABLE 3. Accuracies for unimodal on CREMA-D Dataset.

Model	Accuracy
AlexNet	51.13
ResNet-50	53.08
CNN [39]	55.01
VGG-16	57.23
VGG-19	60.17
MMCFF (Proposed)	**72.39**

TABLE 4. Performance evaluation of the Fusion method (proposed method).

Proposed Multimodal Techniques	Accuracy
Decision Fusion	72.24
Data Fusion	76.10
Feature Fusion	**79.81**

experiment, we achieved an accuracy of 76.10% using the early fusion or data level fusion technique. In the second experiment, we used the middle fusion or feature level fusion technique, which resulted in an accuracy of 79.81%. This was better than the data level fusion and decision level fusion techniques. In the third and final experiment, we utilized the late fusion or decision level fusion technique, which achieved a testing accuracy of 72.24%, as mentioned in Table 4. These experiments demonstrate the effectiveness of our proposed multimodal methodology based on deep learning in recognizing facial expressions under masked conditions.

By applying feature-level fusion, the multimodal has been improved by up to 4% to 6% with other fusion techniques. Looking at the bigger picture, the improvement of a multimodal technique enhances the overall test accuracy from 69.89% to 79.81%, which is a 10% increase in MLF-W-FER and CREMA-D datasets as compared to unimodal techniques.

In our multimodal experiments, we use the confusion matrix to calculate testing accuracies, allowing for a detailed breakdown of predicted and actual labels. It facilitates performance analysis and helps us evaluate accuracy based on the F1 score. These measures are crucial when assessing the effectiveness of classification models. Figure 8 visually represents the confusion matrix from our multimodal experiments.

TABLE 5. Unimodal and multimodal facial expression recognition under mask comparison between proposed and existing technologies.

Architectures	Type	Dataset	Accuracy
AlexNet [1]	Unimodal	MLF-W-FER	55.00
VGG-16 [1]	Unimodal	MLF-W-FER	56.73
ResNET-50 [1]	Unimodal	MLF-W-FER	56.80
VGG-19 [28]	Unimodal	MLF-W-FER	56.79
Mobile Net [28]	Unimodal	MLF-W-FER	66.41
Unimodal (Proposed)	Unimodal	MLF-W-FER	68.76
Decision Fusion (Proposed)	Multimodal	MLF-W-FER & CREMA-D	72.24
MMAFER [39]	Multimodal	MLF-W-FER & CREMA-D	75.67
Data Fusion (Proposed)	Multimodal	MLF-W-FER & CREMA-D	76.10
Feature Fusion (Proposed)	Multimodal	MLF-W-FER & CREMA-D	**79.81**

Recent research has shown a growing interest in studying the impact of masks on facial and voice expressions. For instance, a study by the author [39] utilizing a multimodal approach achieved an accuracy of 75.67. However, our study, which used a refined multimodal fusion technique, yielded a significantly higher accuracy of 79.81. These results highlight the crucial role of considering multiple modalities and utilizing advanced methods in analyzing expression patterns. Ultimately, our findings contribute to a deeper understanding of the complex relationship between masks and voice expressions.

Table 5 provides an overview of the accuracies obtained from the MLF-W-FER unimodal approach and the multimodal approach applied to both the MLF-W-FER and CREMA-D datasets. It is worth noting that when evaluating two different datasets with distinct image characteristics, feature-level fusion demonstrates superior performance.

VI. CONCLUSION

Our paper proposes a multimodal technique to enhance the modal on complex datasets like MLF-W-FER. The approach involves utilizing the voice emotions dataset of CREMA-D to evaluate the masked facial expressions of MLF-W-FER. We also employed various techniques to avoid overfitting and enhance precision. By applying the right combination of regularization and fusion techniques, the feature-level fusion multimodal approach improved accuracy by up to 10% compared to the single model technique. Our experiments with M-LFW-FER and CREMA-D datasets yielded an accuracy of 79.81%, making our proposed multimodal architecture a promising solution for improving accuracy on complicated datasets.

Despite the accuracy mentioned above, the work has some limitations in the various orientation scenarios. In fusion techniques, it is essential to consider that various modalities come with distinct scales, units, or data types. These differences can pose challenges for early fusion methods. If properly preprocessed, these methods might avoid difficulties when combining dissimilar data. Furthermore, it is crucial to realize that the datasets employed for training and assessment might carry biases toward specific demographics, cultural elements, or emotional categorizations.

As we look to the future of facial expression analysis under masks, it is essential to consider the impact of cultural diversity. Different cultures may exhibit unique vocal tones and facial expressions, which can affect the accuracy of our models. To improve the effectiveness of our analysis, we need to incorporate a broader range of expressions - both with and without masks - to achieve more reliable and favorable results. By doing so, we can ensure that our models are more robust and inclusive of all cultures and communities.

REFERENCES

[1] H. M. Shahzad, S. M. Bhatti, A. Jaffar, S. Akram, M. Alhajlah, and A. Mahmood, "Hybrid facial emotion recognition using CNN-based features," *Appl. Sci.*, vol. 13, no. 9, p. 5572, 2023, doi: 10.3390/app13095572.

[2] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *Int. J. Cognit. Comput. Eng.*, vol. 2, pp. 57–64, Jun. 2021, doi: 10.1016/j.ijcce.2021.02.002.

[3] J. Hernandez, J. Lovejoy, D. McDuff, J. Suh, T. O'Brien, A. Sethumadhavan, G. Greene, R. Picard, and M. Czerwinski, "Guidelines for assessing and minimizing risks of emotion recognition applications," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Nara, Japan, 2021, pp. 1–8, doi: 10.1109/ACII52823.2021.9597452.

[4] M. Marini, A. Ansani, F. Paglieri, F. Caruana, and M. Viola, "The impact of facemasks on emotion recognition, trust attribution and re-identification," *Sci. Rep.*, vol. 11, no. 1, p. 5577, Mar. 2021, doi: 10.1038/s41598-021-84806-5.

[5] Y. Kong, Z. Ren, K. Zhang, S. Zhang, Q. Ni, and J. Han, "Lightweight facial expression recognition method based on attention mechanism and key region fusion," *J. Electron. Imag.*, vol. 30, no. 6, Nov. 2021, Art. no. 063002, doi: 10.1117/1.jei.30.6.063002.

[6] F. Grundmann, K. Epstude, and S. Scheibe, "Face masks reduce emotion-recognition accuracy and perceived closeness," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0249792, doi: 10.1371/journal.pone.0249792.

[7] F. Pazhoohi, L. Forby, and A. Kingstone, "Facial masks affect emotion recognition in the general population and individuals with autistic traits," *PLoS ONE*, vol. 16, no. 9, Sep. 2021, Art. no. e0257740, doi: 10.1371/journal.pone.0257740.

[8] T. Puri, M. Soni, G. Dhiman, O. I. Khalaf, M. Alazzam, and I. R. Khan, "Detection of emotion of speech for RAVDESS audio using hybrid convolution neural network," *J. Healthcare Eng.*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/8472947.

[9] M. N. A. Tawhid, S. Siuly, H. Wang, F. Whittaker, K. Wang, and Y. Zhang, "A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from EEG," *PLoS ONE*, vol. 16, no. 6, Jun. 2021, Art. no. e0253094, doi: 10.1371/journal.pone.0253094.

[10] V. Franzoni, G. Biondi, and A. Milani, "Emotional sounds of crowds: Spectrogram-based analysis using deep learning," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 36063–36075, Aug. 2020, doi: 10.1007/s11042-020-09428-x.

[11] M. Grahlow, C. I. Rupp, and B. Derntl, "The impact of face masks on emotion recognition performance and perception of threat," *PLoS ONE*, vol. 17, no. 2, Feb. 2022, Art. no. e0262840, doi: 10.1371/journal.pone.0262840.

[12] N. Mheidly, M. Y. Fares, H. Zalzale, and J. Fares, "Effect of face masks on interpersonal communication during the COVID-19 pandemic," *Frontiers Public Health*, vol. 8, p. 898, Dec. 2020, doi: 10.3389/fpubh.2020.582191.

[13] N. Abbaspoor and H. Hassanpour, "Face recognition in a large dataset using a hierarchical classifier," *Multimedia Tools Appl.*, vol. 81, no. 12, pp. 16477–16495, Mar. 2022, doi: 10.1007/s11042-022-12382-5.

[14] S. Vachmanus, A. A. Ravankar, T. Emaru, and Y. Kobayashi, "Multi-modal sensor fusion-based semantic segmentation for snow driving scenarios," *IEEE Sensors J.*, vol. 21, no. 15, pp. 16839–16851, Aug. 2021, doi: 10.1109/JSEN.2021.3077029.

[15] Q. Abbas, M. E. A. Ibrahim, and M. A. Jaffar, "A comprehensive review of recent advances on deep vision systems," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 39–76, May 2018, doi: 10.1007/s10462-018-9633-3.

[16] J. McAvoy, A. Creed, A. Cotter, L. Merriman, P. O'Reilly, M. Dempsey, and A. Brennan, "Investor decision making: An investigation of the modality effect," *J. Decis. Syst.*, vol. 32, pp. 1–22, Jan. 2022, doi: 10.1080/12460125.2021.2023256.

[17] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, May 2022, doi: 10.1016/j.inffus.2021.12.003.

[18] N. Jaafar and Z. Lachiri, "Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance," *Exp. Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118523, doi: 10.1016/j.eswa.2022.118523.

[19] K. Wang, Y. Song, Z. Huang, Y. Sun, J. Xu, and S. Zhang, "Additive manufacturing energy consumption measurement and prediction in fabricating lattice structure based on recallable multimodal fusion network," *Measurement*, vol. 196, Jun. 2022, Art. no. 111215, doi: 10.1016/j.measurement.2022.111215.

[20] W. Sun, X. Chen, X. Zhang, G. Dai, P. Chang, and X. He, "A multi-feature learning model with enhanced local attention for vehicle re-identification," *Comput., Mater. Continua*, vol. 69, no. 3, pp. 3549–3561, 2021, doi: 10.32604/cmc.2021.021627.

[21] A. S. Al-Waisy, R. Qahwaji, S. Ipson, and S. Al-Fahdawi, "A multimodal deep learning framework using local feature representations for face recognition," *Mach. Vis. Appl.*, vol. 29, no. 1, pp. 35–54, Sep. 2017, doi: 10.1007/s00138-017-0870-2.

[22] W. Wei, Q. Jia, Y. Feng, G. Chen, and M. Chu, "Multi-modal facial expression feature based on deep-neural networks," *J. Multimodal User Interfaces*, vol. 14, no. 1, pp. 17–23, Jul. 2019, doi: 10.1007/s12193-019-00308-9.

[23] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–8, doi: 10.1109/ijcnn.2015.7280539.

[24] S. A. Kashinath, S. A. Mostafa, A. Mustapha, H. Mahdin, D. Lim, M. A. Mahmoud, M. A. Mohammed, B. A. S. Al-Rimy, M. F. M. Fudzee, and T. J. Yang, "Review of data fusion methods for real-time and multi-sensor traffic flow analysis," *IEEE Access*, vol. 9, pp. 51258–51276, 2021, doi: 10.1109/ACCESS.2021.3069770.

[25] A. Gumaei, W. N. Ismail, M. R. Hassan, M. M. Hassan, E. Mohamed, A. Alelaiwi, and G. Fortino, "A decision-level fusion method for COVID-19 patient health prediction," *Big Data Res.*, vol. 27, Feb. 2022, Art. no. 100287, doi: 10.1016/j.bdr.2021.100287.

[26] K. Jin, Y. Yan, M. Chen, J. Wang, X. Pan, X. Liu, M. Liu, L. Lou, Y. Wang, and J. Ye, "Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration," *Acta Ophthalmologica*, vol. 100, no. 2, pp. 512–520, Jun. 2021, doi: 10.1111/aos.14928.

[27] W. Nsengiyumva, S. Zhong, M. Luo, Q. Zhang, and J. Lin, "Critical insights into the state-of-the-art NDE data fusion techniques for the inspection of structural systems," *Structural Control Health Monitor.*, vol. 29, no. 1, p. e2857, Sep. 2021, doi: 10.1002/stc.2857.

[28] B. Yang, J. Wu, and G. Hattori, "Facial expression recognition with the advent of face masks," in *Proc. 19th Int. Conf. Mobile Ubiquitous Multimedia*, Nov. 2020, pp. 335–337, doi: 10.1145/3428361.3432075.

[29] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014, doi: 10.1109/TAFFC.2014.2336244.

[30] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7169–7173, doi: 10.1109/ICASSP40776.2020.9054317.

[31] S. A. Gebereselassie and B. K. Roy, "Secure speech communication based on the combination of chaotic oscillator and logistic map," *Multimedia Tools Appl.*, vol. 81, no. 18, pp. 26061–26079, Mar. 2022, doi: 10.1007/s11042-022-12803-5.

[32] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5, doi: 10.1109/PlatCon.2017.7883728.

[33] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6669–6673, doi: 10.1109/ICASSP.2013.6638952.

[34] A. Poulose, C. S. Reddy, J. H. Kim, and D. S. Han, "Foreground extraction based facial emotion recognition using deep learning xception model," in *Proc. 12th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Aug. 2021, pp. 356–360, doi: 10.1109/ICUFN49451.2021.9528706.

[35] F. Liu, H. Xu, M. Qi, D. Liu, J. Wang, and J. Kong, "Depth-wise separable convolution attention module for garbage image classification," *Sustainability*, vol. 14, no. 5, p. 3099, Mar. 2022, doi: 10.3390/su14053099.

[36] R. Raza, F. Zulfiqar, S. Tariq, G. B. Anwar, A. B. Sargano, and Z. Habib, "Melanoma classification from dermoscopy images using ensemble of convolutional neural networks," *Mathematics*, vol. 10, no. 1, p. 26, Dec. 2021, doi: 10.3390/math10010026.

[37] L. Qian, L. Hu, L. Zhao, T. Wang, and R. Jiang, "Sequence-dropout block for reducing overfitting problem in image classification," *IEEE Access*, vol. 8, pp. 62830–62840, 2020, doi: 10.1109/access.2020.2983774.

[38] L. Chen, M. Li, X. Lai, K. Hirota, and W. Pedrycz, "CNN-based broad learning with efficient incremental reconstruction model for facial emotion recognition," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10236–10241, 2020, doi: 10.1016/j.ifacol.2020.12.2754.

[39] H. M. Shahzad, S. M. Bhatti, A. Jaffar, and M. Rashid, "A multi-modal deep learning approach for emotion recognition," *Intell. Autom. Soft Comput.*, vol. 36, no. 2, pp. 1561–1570, 2023, doi: 10.32604/iasc.2023.032525.

H. M. SHAHZAD received the master's degree from the COMSATS Institute of Information Technology, Lahore, Pakistan. He is currently pursuing the Ph.D. degree with Superior University, Lahore.

He is an Assistant Professor with Superior University. His current research interests include image processing, machine learning, deep learning, and computing vision.

SOHAIL MASOOD BHATTI received the Ph.D. degree in computer science from the National University of Computer and Emerging Sciences, Islamabad, in 2014.

He was a Postdoctoral Researcher with the Department of Electrical Engineering, Universidad De Chile, Santiago, Chile, in 2015. He has over ten years of experience in machine learning and over ten years of experience in the software industry. He has a U.S. patent and over 20 research papers in his credit. He is interested in conducting research in areas related to image/signal processing, applied machine learning, artificial intelligence, malware detection, and medical image processing. He has special interest in machine learning-based techniques and their applications. He has been involved in developing techniques for image restoration, malware detection, and functions optimization problems. He has also worked on different real-world machine learning applications in heart disease detection, cancer grading, volcanology, and mining using signal processing techniques as well as on deep neural networks.

ARFAN JAFFAR received the M.Sc. degree in computer science from Quaid-i-Azam University Islamabad, Islamabad, Pakistan, in March 2003, and the M.S. and Ph.D. degrees in computer science from the FAST National University of Computer and Emerging Sciences, in 2006 and 2009, respectively.

He received a postdoctoral research fellowship from South Korea and carried-out research at the top raking Korean university Gwangju Institute of Science and Technology, Gwangju, South Korea, from 2010 to 2013. He is a Reviewer of 30 reputed international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, *Pattern Recognition*, *Knowledge*, *Information Sciences*, and *Information*.

MUHAMMAD RASHID received the Ph.D. degree in computer science from the National University of Computer and Emerging Sciences (NUCES), Pakistan. He was an Assistant Professor with Foundation University, Islamabad, Pakistan, from 2010 to March 2012. He was the Head of the Department of IT, Rustaq College of Applied Sciences, Oman, from 2012 to August 2018. He is currently the Head of the Department of Computer Science, National University of Technology. His research interests include software development, programming, and artificial intelligence.

SHEERAZ AKRAM received the Master of Science degree in computer science from the Lahore University of Management Sciences (LUMS), Lahore, Pakistan, and the Ph.D. degree in software engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan. He completed his postdoctoral research training from the University of Pittsburgh, USA, and worked on a project funded through grant U01 HL137159. He has experience of 17 years of working at universities which includes three years of international research experience. He is currently with the Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. He is also associated with the Department of Computer Science, Faculty of Computer Science and Information Technology, Superior University, Lahore. His research interests include data science, medical image processing, artificial intelligence in data science, machine learning, deep learning, computer vision, and digital image processing.